Deep Learning & Optimisation – Prediction Task

1. Description

Deep learning is widely used to support medical diagnoses such as cancer prediction, diabetes prediction and x-ray categorisation. This challenge is designed to predict the possibility to have diabetes based on past medical records.

You are given a data set that contains patient medical records of Pima Indians and whether they had an onset of diabetes within five years. Using these data you need to build and optimise a **neural network** based on the concepts we covered during the sessions (or your own research on the covered topics) to predict the possibility to have diabetes.

Note: Use 100 as the value for any random state you set to generate comparative results.

2. Data set

Within the data folder, there are 2 .csv files which are named “train.csv” and “test.csv”. The details of the attributes mentioned in both files are as follows.

- A1 - Number of times pregnant
- A2 - Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- A3 - Diastolic blood pressure (mm Hg)
- A4 - Triceps skin fold thickness (mm)
- A5 - 2-Hour serum insulin (mu U/ml)
- A6 - Body mass index (weight in kg/(height in m)^2)
- A7 - Diabetes pedigree function
- A8 - Age (years)

train.csv
Training data to use with model training and validation.
Number of entries: 668
Columns:

- id - Unique Id assigned to each person in the training data set
- attributes - A1 – A8
- class - 0 or 1 (1= tested positive for diabetes)

test.csv
Testing data to use with predictions.
Number of entries: 100
Columns:

- id - Unique Id assigned to each person in the test data set
- attributes - A1 – A8

3. Submission

The final submission should include the following:

a. Test data predictions

Your test predictions need to be submitted to the **CodaLab**. Please carefully read the submission instructions given on the CodaLab page before submission.

Make sure you use the **best** model(s) you built to make predictions on the test dataset.

b. Blog explaining your implementation

A maximum 300-word blog-style document explaining what you did and why including any plots and screenshots of the code needs to be submitted to **Moodle**.

Make sure to include evaluation results that you use to compare the models you built to select the best model in your blog.

Note: Both elements mentioned above **must be** submitted to consider your submission as a valid submission.

Hints:

Model evaluation

Use 70% of the training data to train the machine learning model and the remaining 30% as the validation set to evaluate the model.

The following code segment can be used to perform data splitting.

```
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.3,
random_state=100)
```

Save Predictions

The following code allows saving the predictions into a .csv file. Make sure to fill in the missing parts in this code according to your training process and use the correct model to make predictions.

```
# read csv file into a dataframe
df_test = pd.read_csv('/content/test.csv')

# extract features from test data set
X_test = <add your code here>

# make predictions using the best model
predictions = model.predict(X_test)

# convert predictions to the required label format (0 or 1)
final_predictions = <add your code here>

# create data frame for submission
df_test = pd.DataFrame(df_test['id'])
df_test['prediction'] = final_predictions

# save data frame to .csv file
df_test.to_csv('/content/test-predictions.csv', index=False)
```

The saved .csv file can be used to generate the final submission file using the Python scripts provided with the CodaLab page.